Multimedijalne baze podataka u upravljanju nematerijalnim kulturnim nasleđem

Ivana Tomašević

JeRTeh seminar, Društvo za jezičke resurse i tehnologije

oktobar, 2021.

Sadržaj

Motivacija

- Istraživanje sprovedeno motivisano je multimedijalnom kolekcijom koja je predstavljala rezultat terenskog istraživanja koje su izveli istraživači Balkanološkog instituta SANU
- Istraživači su intervjuisali lokalno stanovništvo na različitim lokacijama u području Balkana
- Kolekcija se sastoji od audio i video materijala, fotografija, rukopisa i tekstualnih opisa - protokola
- Ne postoji uspostavljena organizacija koja bi omogućila efikasno pretraživanje date kolekcije

Pregled oblasti

- Upotrebom informacionih tehnologija i digitalizacije kao rezultat u domenu kulturnog nasleđa je stvaranje digitalnih baza podataka o nematerijalnoj kulturnoj baštini
- CIDOC Conceptual Reference Model (CRM) je formalna ontologija za transformaciju i integraciju informacija iz domena kulturnog nasleđa
- Dublin Core inicijativa za metapodatke je veoma korišćen standard za definisanje metapodataka kojima se opisuju kulturna dobra

Pregled oblasti

- Multimedijalna baza podataka je kolekcija digitalnih dokumenata koji mogu biti tekst, grafika (crteži i ilustracije), slika, animacija, audio ili video materijali
- Za efikasno rukovanje ovakvim raznovrsnim i složenim podacima neophodan je sistem za upravljanje prilagođen specifičnim zahtevima
- Postojeće baze podataka kulturnog nasleđa koriste pretrage različitih vrsta medija
- Pretraga teksta na prirodnom jeziku je nedovoljno zastupljena u ovakvim bazama, a pogotovo u kontekstu kulturnog nasleđa Balkana

Problem upravljanja nematerijalnim kulturnim nasleđem

- Zahtevi koji se postavljaju pred odgovarajuće sisteme
  - korisničke potrebe i funkcionalnosti
  - sistemski zahtevi
  - izbor tehnologije
- Izazovi u obradi tekstova iz domena kulturnog nasleđa
  - jezik nestandardan ili arhaičan
  - nedovoljni resursi i alati prilagođeni domenu
  - različiti standardi
- Specifičnosti kulturnog nasleđa na prostoru Balkana
  - nedovoljno istražen domen u kontekstu računarske obrade prirodnog jezika
  - specifična pravila i ontologije, specifični rečnici, semantičke strukture

Zadaci koji se rešavaju u okviru ovog istraživanja

- Razvoj adekvatnog dizajna i implementacije multimedijalne baze podataka nematerijalnog kulturnog nasleđa koja bi odgovarala potrebama različitih korisnika
- Automatska semantička anotacija protokola metodama obrade prirodnog jezika kao osnova za polu-automatsku anotaciju multimedijalne kolekcije i uspešnu tematsku pretragu i pretragu po metapodacima koji su u skladu sa CIDOC CRM standardom
- Povezivanje sadržaja baze sa geolokacijskim informacijama na mapi
- Istraživanje dodatnih mogućnosti pretrage ove kolekcije u cilju dobijanja novih znanja
- Razvoj opšteg metodološkog okvira za rešavanje sličnih zadataka i u drugim domenima

Tehnike koje su korišćene za rešavanje problema upravljanja nematerijalnim kulturnim nasleđem

- Ekstrakcija informacija iz teksta
- Klasifikacija teksta
- Anotacija dokumenata
- Pretraživanje dokumenata

Ekstrakcija informacija iz teksta

- Korišćene su metode zasnovane na pravilima za ekstrakciju:
  - imenovanih entiteta
  - tema

Ekstrakcija imenovanih entiteta

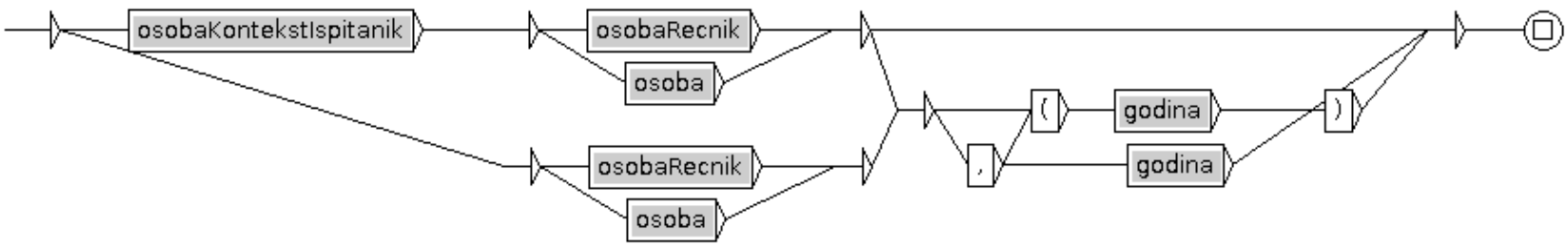
- Za ekstrakciju imenovanih entiteta koriste se:
  - elektronski rečnici
    - na primer, reč "Beograd" se prepoznaje kao lokacija ako se nalazi u rečniku toponima
  - kontekst
    - na primer, "živi u X", X se prepoznaje kao lokacija, čak i ako X nije nađeno u odgovarajućem rečniku
  - kontekst sa posebnim potkontekstom ("subNE") opšteg NE, koji je određen specifičnim zahtevima domena
    - na primer, "informator je A" - A se prepoznaje kao potklasa "ispitanik" klase "osoba", ili "razgovor vodila A" - A se prepoznaje kao potklasa "ispitivač" klase "osoba"

Ekstrakcija imenovanih entiteta

- Lista obrađenih imenovanih entiteta:
  - Oznaka
  - Ispitanik
  - Ispitivač
  - Osoba-ostali
  - Lokacija
  - Datum
  - Godina
  - Jezik
  - Etnicitet
  - Religija

Ekstrakcija imenovanih entiteta

- Za ekstrahovanje imenovanih entiteta razvijen je 29 pomoćnih transduktora
- Na primer, za ekstrahovanje imenovanog entiteta tipa osoba, podtipa ispitanik razvijeni su transduktori

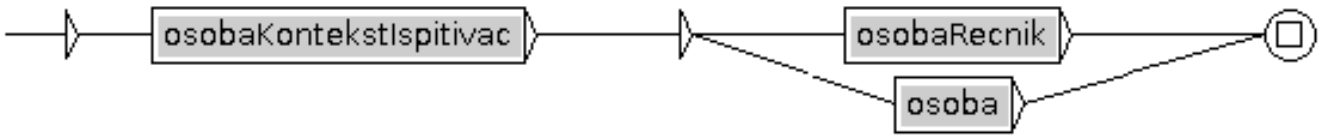


Slika: Transduktor "Ispitanik"

- Primeri ekstrahovanja entiteta ispitanik su: "Sara Petrović (1950)" ili "razgovor je vođen sa Sarom Petrović" - "razgovor je vođen sa" prepoznaje se kao "osobaKontekstIspitanik", dok se "Sarom Petrović" prepoznaje na osnovu konteksta bez obzira da li postoji u rečniku ličnih imena

Ekstrakcija imenovanih entiteta

- Na sličan način se ekstrahuje i entitet klase Ispitivač



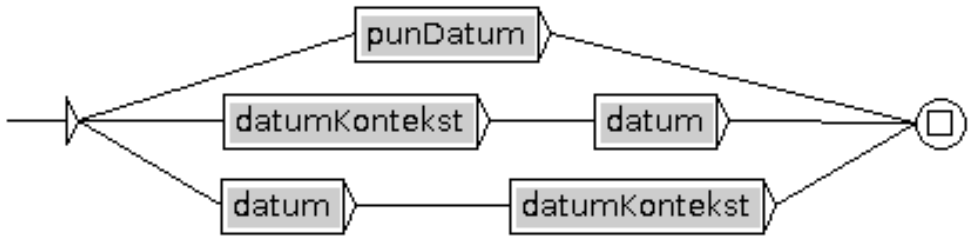
Slika: Transduktor "Ispitivac"

- Entitet klase Osoba-ostali se ekstrahuje samo na osnovu rečnika

Slika: Transduktor "Osoba-ostali"

Ekstrakcija imenovanih entiteta

- Na sličan način se prepoznaju i ostali imenovani entiteti



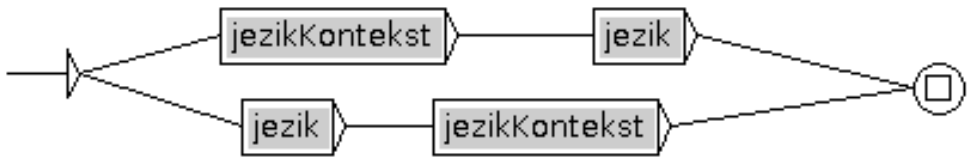
Slika: Transduktor "Datum"

Slika: Transduktor "Godina"

Slika: Transduktor "Lokacija"

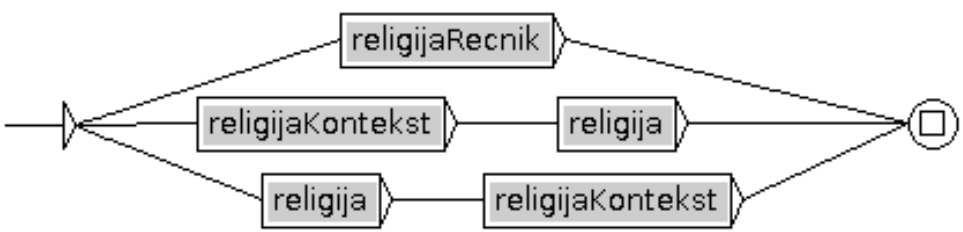
Ekstrakcija imenovanih entiteta

- Na sličan način se prepoznaju i ostali imenovani entiteti



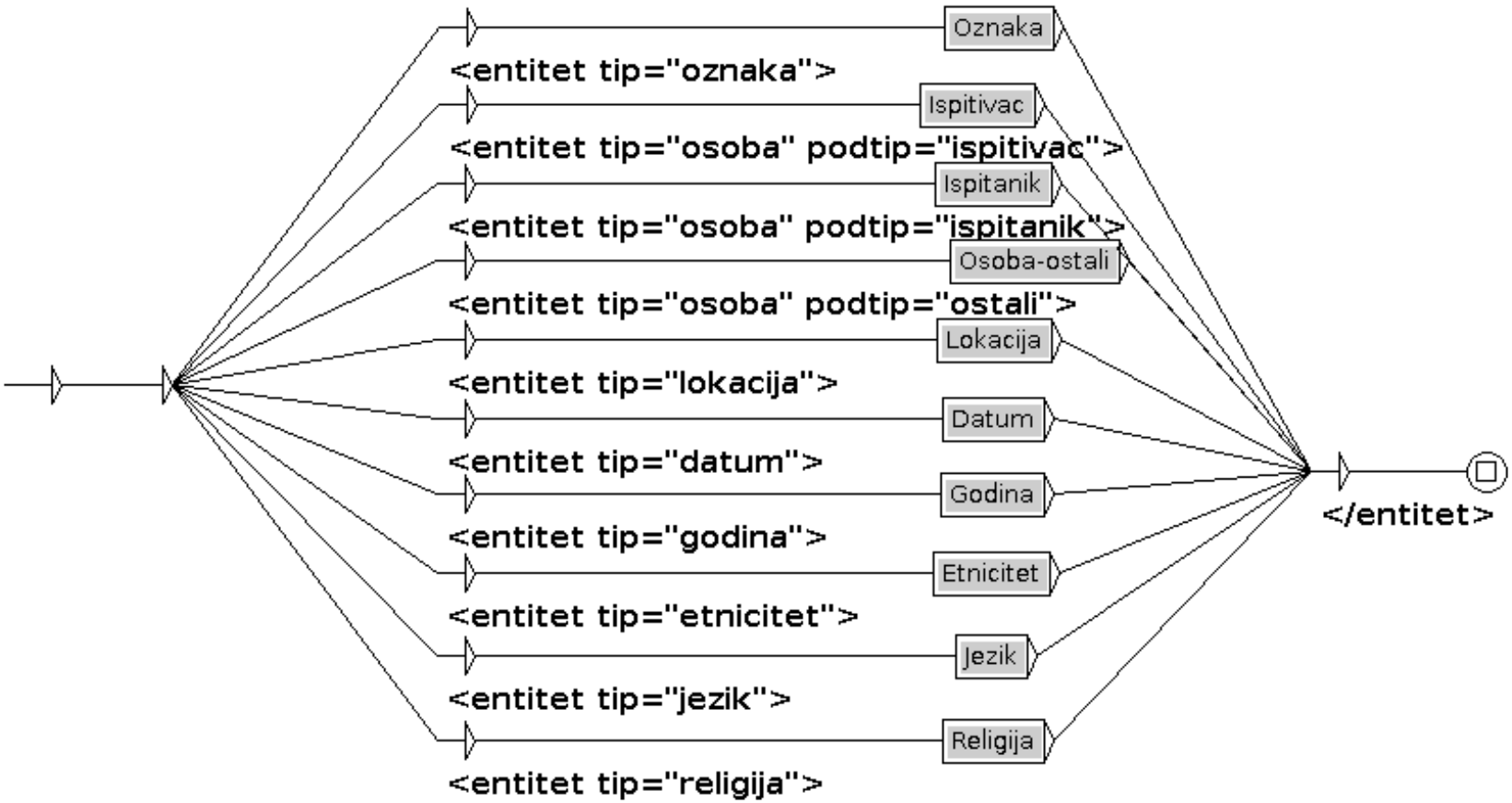
Slika: Transduktor "Jezik"

Slika: Transduktor "Etnicitet"



Slika: Transduktor "Religija"

Ekstrakcija imenovanih entiteta



Slika: Transduktor za ekstrakciju imenovanih entiteta

Ekstrakcija tema

- Ekstrakcija tema zasnovana je na konstrukciji semantičkih struktura:
  - uočavaju se kontekstne fraze koje označavaju izabrane teme
    - na primer, “sejalo se žito”
  - identifikuju se semantičke strukture koje opisuju uočene kontekstne fraze
    - na primer, “biljkePoslovi” + “pomoćne reči” + “biljkeTermi”
  - identifikuju se klase termova koje sačinjavaju semantičke strukture, kao i liste takvih termova
    - na primer, klasa “biljkePoslovi” sadrži reči “sejanje”, “uzgoj”, “branje”; klasa “biljkeTermi” sadrži reči poput “žito”, “kukuruz”, “krompir”

Ekstrakcija tema

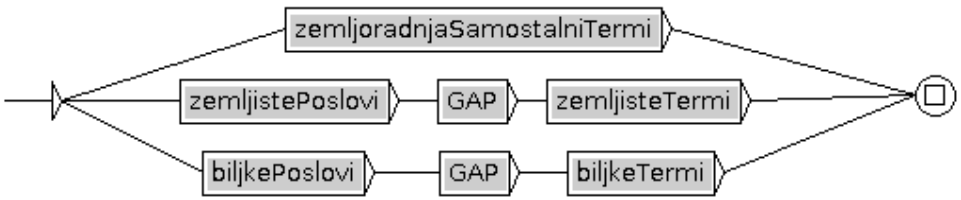
- Semantičke strukture koje odgovaraju frazama iz određene tematike se tada mogu podeliti na:
  - samostalne terme
    - na primer, “poljoprivreda”, “stočarstvo”, “zemljoradnja”
  - specifične terme koji se nalaze u specifičnom kontekstu
    - na primer, u frazi “sejalo se žito”, “žito” je specifični term klase “biljkeTermi”, dok je “sejalo” specifični kontekst klase “biljkePoslovi”, pri čemu se dozvoljava da se između nađe određeni broj pomoćnih reči koji će se u nastavku nazivati GAP
  - specifične terme koji se nalaze u opštem kontekstu
    - na primer, u frazi “terminologija sejača”, “terminologija” je term iz klase “opštiKontekst”, dok je “sejača” specifični term klase termova “poljoprivredaAlati”

Ekstrakcija tema

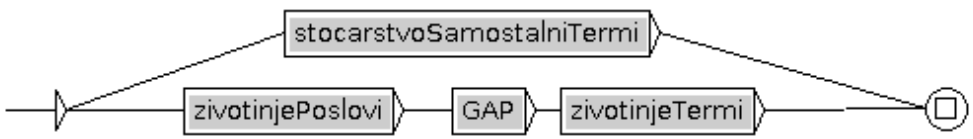
- Obrađene teme iz tematike narodna privreda:
  - Domaća radinost
  - Lov i ribolov
  - Pčelarstvo
  - Poljoprivreda
  - Rudarstvo
  - Šumarstvo
  - Trgovina
  - Zanatstvo

Ekstrakcija tema

- Za prepoznavanje i ekstrahovanje tema razvijen je 41 pomoćni transduktor
- Primer: za prepoznavanje fraze iz teme poljoprivreda, pomoćni transduktori su



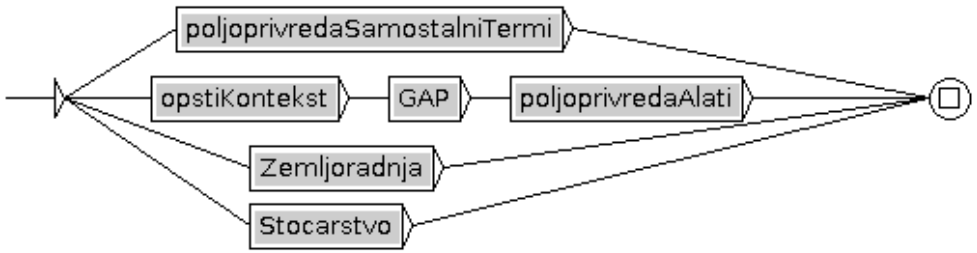
Slika: Transduktor "Zemljoradnja"



Slika: Transduktor "Stocarstvo"

Ekstrakcija tema

- Koristeći prethodne, ali i dodavanjem novih transduktora, dobija se transduktor "Poljoprivreda"

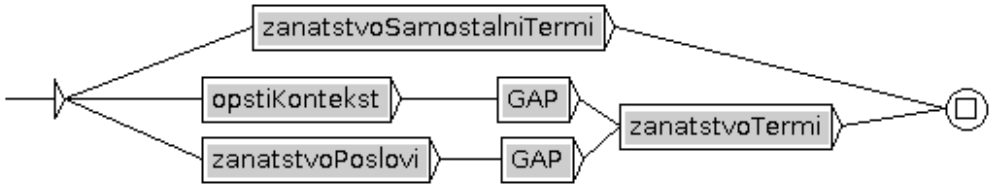


Slika: Transduktor "Poljoprivreda"

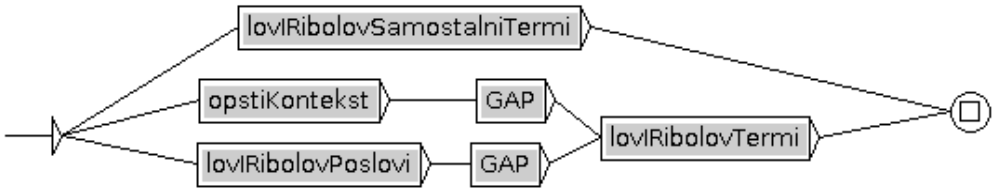
- Primeri prepoznavanja fraze na osnovu strukture "biljkePoslovi" + GAP + "biljkeTermi" su "sejali su žito" ili "sejanje žita"

Ekstrakcija tema

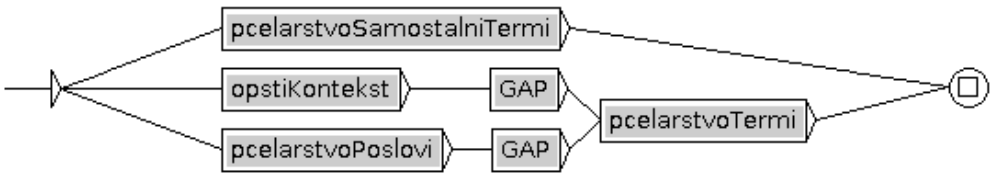
- Na sličan način prepoznaju se i ostale teme



Slika: Transduktor "Zanatstvo"



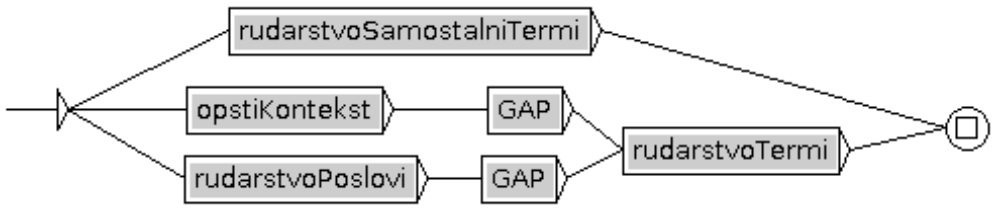
Slika: Transduktor "LovIRibolov"



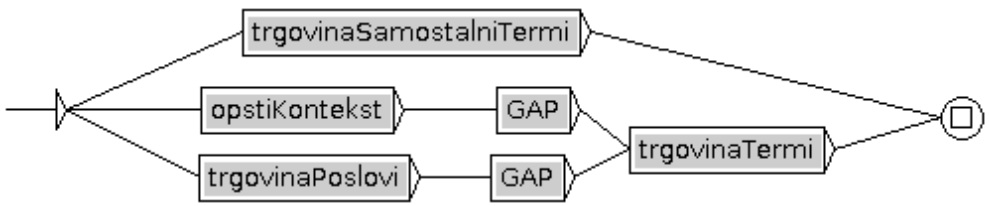
Slika: Transduktor "Pcelarstvo"

Ekstrakcija tema

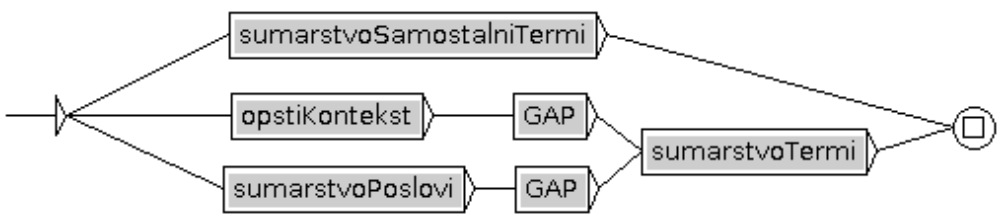
- Na sličan način prepoznaju se i ostale teme



Slika: Transduktor "Rudarstvo"

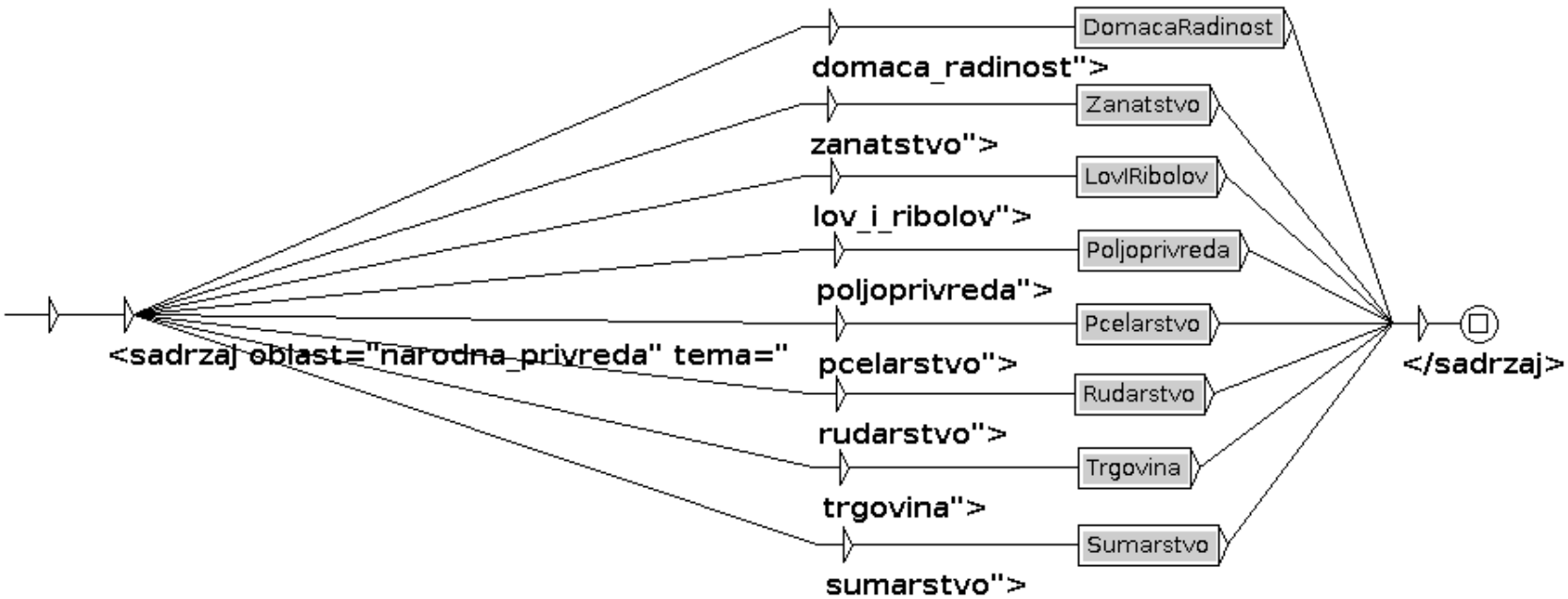


Slika: Transduktor "Trgovina"



Slika: Transduktor "Sumarstvo"

Ekstrakcija tema



Slika: Transduktor za ekstrakciju tema iz oblasti “narodna privreda”

Klasifikacija teksta

- Izvršeno je poređenje kvaliteta klasifikacije za različite reprezentacije teksta i metode klasifikacije
  - Sprovedena su istraživanja kojima je ispitivano kako različiti tipovi *n*-grama (bajt, karakter, reč) u reprezentaciji teksta, i različite metode (SVM, kNN, MaxEnt) utiču na kvalitet klasifikacije različitih skupova tekstova u istom domenu (filmske recenzije) na različitim jezicima (engleski, španski, arapski, francuski, češki, turski i srpski)
- Primenjena je izabrana metoda na klasifikaciju protokola uz poboljšanje semantičkim atributima. Izvršena je klasifikacija teksta pri predstavljanju dokumenata:
  - *n*-gramima karaktera
  - semantičkim atributima
  - hibridnom metodom

Klasifikacija teksta

- Semantički atribut je bilo kakva jedinica koja se može nedvosmisleno definisati, a koja ima neko značenje ili strukturu u datoj kolekciji ili domenu
- Primeri semantičkih atributa koji se mogu naći u protokolima:
  - nabrajanje tema uz korišćenje samostalnog terma
    - 1. vršidba
    - 2. stočarstvo
  - kratka linija određene strukture u kojoj se navodi specifični term
    - Tema: pletenje
  - pojavljivanje fraze koja se može opisati nekom od uvedenih semantičkih struktura
    - sejali su žito

Anotacija dokumenata i organizovanje multimedijalne kolekcije u bazu podataka

- Metodološki okvir semantičke ručne anotacije i organizovanja multimedijalne kolekcije u bazu podataka:
  - Izbor liste atributa za svaki od tipova materijala, uz mogućnost dodeljivanja atributa na nivou segmenata, ali i na nivou celog materijala
  - Mogućnost ručne anotacije materijala kroz grafički korisnički interfejs
  - Podela korisnika na grupe na osnovu tipa aktivnosti - unos, revizija ili pregled
  - Interaktivni rad sa prostornim podacima

Ivana Tanasijević, Biljana Sikimić, Gordana Pavlović-Lažetić, "Multimedia Database of the Cultural Heritage of the Balkans", Language Resourses and Evaluation Conference (LREC), European Language Resources Association (ELRA), Istanbul, ISBN 978-2-9517408-7-7, pp 2874-2881, 2012

Anotacija dokumenata i organizovanje multimedijalne kolekcije u bazu podataka

- Metodološki okvir polu-automatizovane anotacije:
  - Upotreba ekstrakcije informacija iz protokola za dobijanje termova po kojima se može vršiti anotacija multimedijalnih materijala

  Ivana Tanasijević, "Toward automatic annotation of cultural heritage documents", IPSI Transactions on Advanced Research, TAR, IPSI, 15(1), 2019
- Upotreba klasifikacije protokola za pridruživanje tema protokolima, a samim tim i materijalima, koji još nisu ručno anotirani

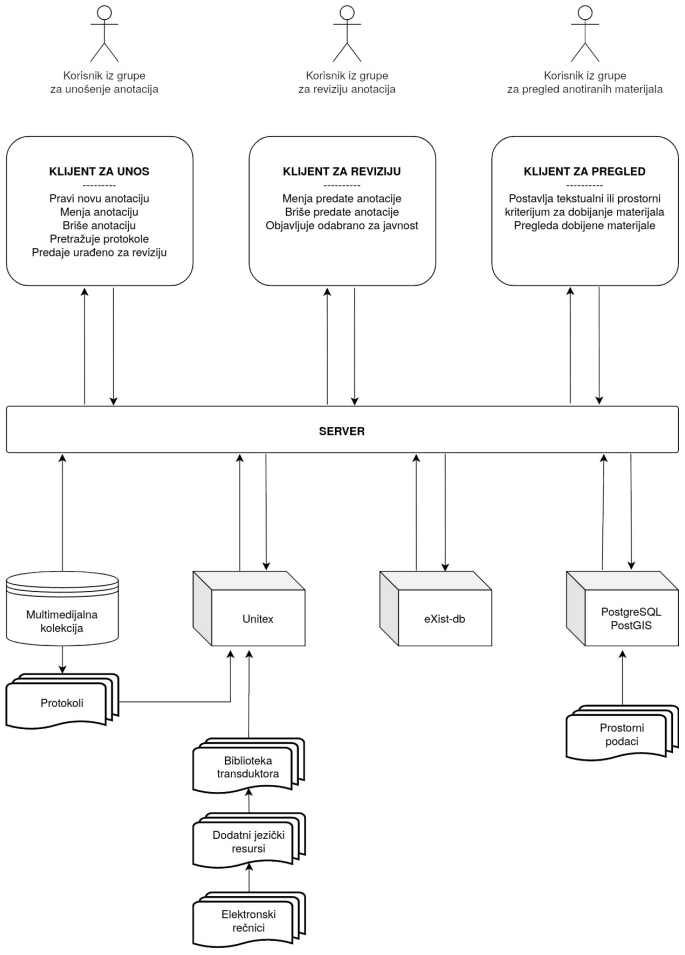
Pretraga multimedijalne baze podataka

- U uspostavljenoj organizaciji pretraga se može vršiti na više načina:
  - pretraga protokola po atributima koji su im pridruženi u fazi automatske semantičke anotacije
  - pretraga materijala po atributima koji su im pridruženi u fazi ručne semantičke anotacije
  - pretraga materijala na osnovu sprovedene klasifikacije protokola po tematici
  - pretraga po prostornom kriterijumu izborom lokacije na interaktivnoj mapi geografskih lokacija

Izbor alata i tehnologija

- eXist baza podataka za rad sa tekstualnim podacima
- PostgreSQL i PostGIS za rad sa prostornim podacima
- Unitex za analize teksta na prirodnom jeziku
- Programski jezik Java za implementaciju sistema u klijent / server organizaciji
- Programski jezik Python za primenu algoritama klasifikacije
- JavaFX za prikaz multimedije

Arhitektura i implementacija



Slika: Arhitektura sistema

Rezultati - Ekstrakcija informacija iz tekstualnih protokola

- Ivana Tanasijević, Gordana Pavlović-Lažetić, "HerCulB: Content-based Information Extraction and Retrieval for Cultural Heritage of the Balkans", The Electronic Library, Emerald Publishing, DOI (10.1108/EL-03-2020-0052), ISSN 0264-0473, 2020 (**M23**)
- Rezultati automatskog indeksiranja dokumenata pokazuju da se ručno pisanim pravilima za ekstrakciju informacija mogu proizvesti visoko kvalitetne semantičke anotacije po izabranoj shemi metapodatka, koja je u skladu sa CIDOC CRM standardom, i sa rečničkim resursima prilagođenim domenu nematerijalnog kulturnog nasleđa

Rezultati - Ekstrakcija informacija iz tekstualnih protokola

- Ukupna F mera za NER koja je dobijena u ovom istraživanju (0.87) u rangu je sa, najsličnijim po domenu, NER sistemima za engleski jezik koji se bave arheološkim domenom
- Za ekstrakciju metapodataka koji se odnose na temu, ukupna F mera (0.90) može se porediti sa sličnim inicijativama koje koriste semantičke strukture, na primer, ekstrakcijom podataka iz narativa na engleskom jeziku medicinskog domena u kojima su dobijeni preciznost 0.93 i odziv 0.83, ili semantičkom anotacijom televizijskih i radio novosti na engleskom jeziku, za koju je objavljena preciznost 1 uz odziv 0.40

Rezultati - Poređenje kvaliteta klasifikacije za različite reprezentacije teksta i metode klasifikacije

- Jelena Graovac, Miljana Mladenović, Ivana Tanasijević, "NgramSPD: Exploring Optimal N-gram Model for Sentiment Polarity Detection in Different Languages", Intelligent Data Analysis 23(2), 2019 (**M23**)
- Na skupovima filmskih recenzija pri detekciji polariteta sentimenata u cilju poređenja različitih reprezentacija dokumenata i metoda klasifikacije, pokazuje se da bajt i karakter *n*-gram modeli daju bolje rezultate od modela koji koriste *n*-grame reči
- Metoda potpornih vektora proizvela je najbolje rezultate u pogledu odziva, i to za *n*-grame karaktera, u kom slučaju je dostignuta F mera od 0.83

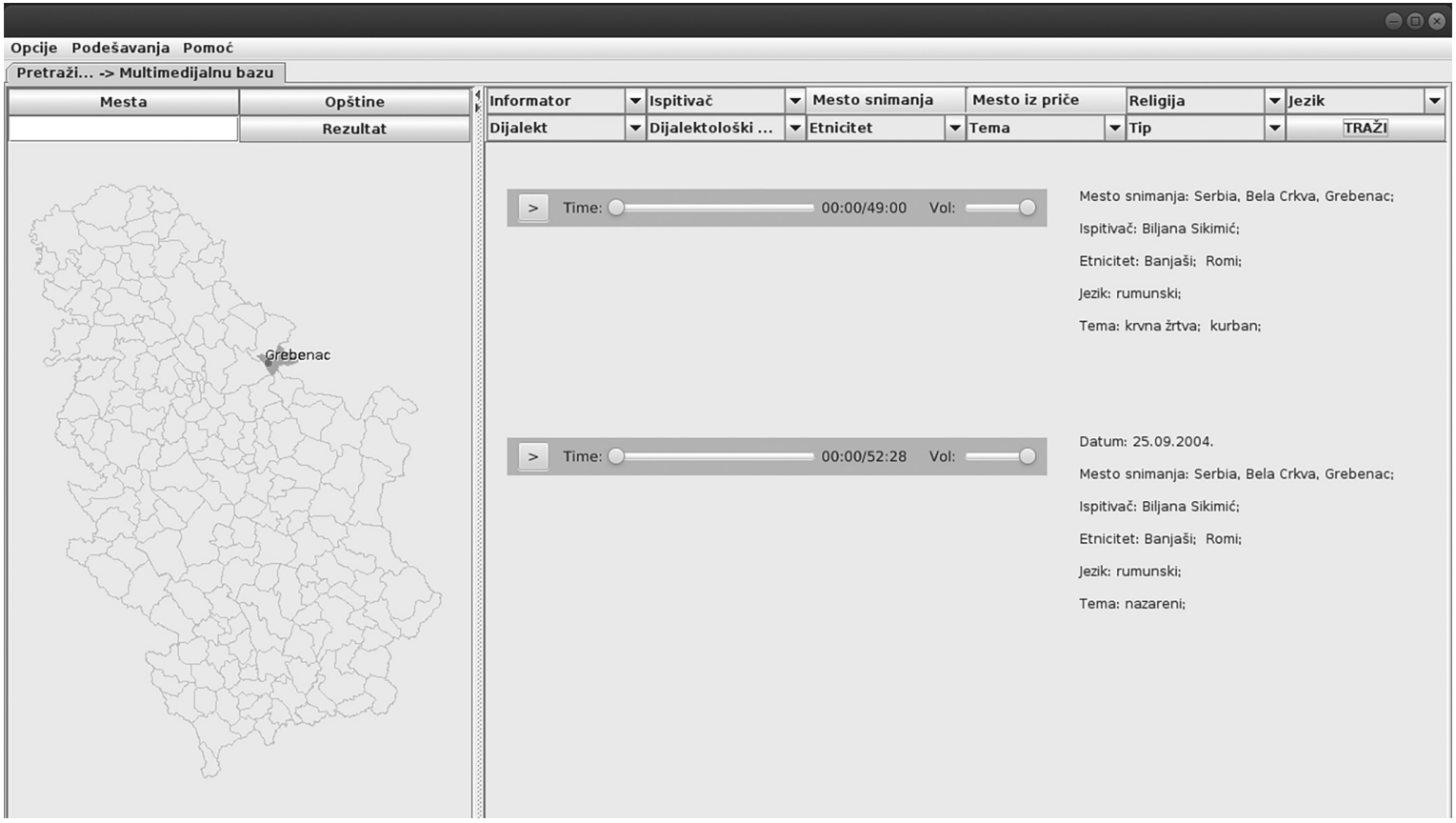
Rezultati - Primena izabranih metoda na zadatak klasifikacije tekstualnih protokola prema tematici

- Metoda potpornih vektora, modifikovana semantičkim atributima, uz predstavljanje dokumenata atributima *n*-grama karaktera postiže F meru od 0.88 u klasifikaciji tekstualnih protokola u odnosu na tematiku narodna privreda

Naučni doprinos

- Razvoj novog modela i implementacija multimedijalne baze podataka nematerijalnog kulturnog nasleđa Balkana
- Razvoj informatičkog modela dokumenata i prostornog modela geografskih karakteristika sadržaja multimedijalne baze podataka nematerijalnog kulturnog nasleđa
- Automatska semantička anotacija sadržaja multimedijalne kolekcije nematerijalnog kulturnog nasleđa, primenom razvijenih metoda ekstrakcije informacija i klasifikacije teksta, kao osnova za uspešnu pretragu po metapodacima
- Razvoj metoda pretrage sadržaja multimedijalne baze podataka nematerijalnog kulturnog nasleđa po prostornim karakteristikama izborom lokacije na interaktivnoj mapi

Naučni doprinos



Slika: Pretraga baze po izabranom kriterijumu i prikaz materijala

Moguća uopštenja

- Razvijena metoda upravljanja multimedijalnom kolekcijom nematerijalnog kulturnog nasleđa Balkana može se primeniti i za organizaciju drugih kolekcija sličnih potreba
- Razvijene metode ekstrakcije informacija i klasifikacije teksta mogu se primeniti u drugim domenima
- Metode obrade prirodnog jezika mogu se uz određena prilagođavanja primeniti i na druge morfološki bogate jezike

Diskusija

- Rezultati eksperimenata pokazuju da korišćenje pristupa zasnovanog na pravilima za zadatak ekstrakcije informacija iz tekstova na prirodnom jeziku, u kombinaciji sa dodatnim jezičkim resursima i uz ulaganje razumnog truda daje veoma dobre rezultate.
- Rezultati eksperimenata klasifikacije teksta ukazuju na zaključak da primenjene semantičke tehnike daju značajan doprinos kvalitetu klasifikacije statističkim metodama mašinskog učenja.
- Pokazuje se da kontekst igra veliku ulogu u zadacima ekstrakcije informacija i klasifikacije teksta

Dalji rad

- Razvijene metode nisu iscrpne u mogućnostima, stoga postoji prostor za dalje unapređenje
- Domen kulturnog nasleđa je veoma bogat semantikom, bilo bi korisno budući rad usmeriti ka istraživanju načina kako da se raznovrsni podaci analiziraju i prikažu u cilju boljeg uvida u stvarnost koja je zapisana njima, ali i otkrivanja novih znanja i dodatnih veza između podataka
- U saradnji sa korisnicima sistema potrebno je definisati nove zahteve koji bi bili nadogradnja postojećim početnim zahtevima
- Složene su potrebe različitih grupa korisnika što usložnjava i zadatak organizacije i upravljanja multimedijalnom kolekcijom
- Za dalji rad na ovom problemu se preporučuje veća uključenost eksperata iz više domena, kako društvenih i humanističkih tako i računarskih nauka

Dalji rad

- Poboljšanje sistema se može videti u bogatijoj vizuelizaciji podataka i interaktivnosti celog sistema sa različitim grupama korisnika, poput studenata, naučnika, eksperata iz različitih oblasti, ali i šire javnosti
- Korisno bi bilo razviti dodatne metode pretrage informacija koje su dobijene uspostavljenom bogatom semantičkom shemom metapodataka
- Zanimljiva tema za buduća istraživanja bi mogla biti povezivanje više aspekata podataka, poput prostora, vremena i drugih karakteristika koje se prostiru kroz prostor i vreme
- Korisno bi bilo povećati skup tekstova i razviti model koji bi mogao da iskoristi prednosti tehnika neuronskih mreža u obradi prirodnih jezika i za druge, složenije analize teksta